International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Deep Learning-Based Protein Family Classification using Recurrent Neural Networks

[1]Dr. M. Muni Babu, [2]M. Mounika, [3]K. Gayathri, [4]N. Vinuthna, [5]T. Gayathri

[1]Associate Professor, Department of Computer Science and Engineering (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

[2]Assistant Professor, Department of Computer Science and Engineering (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

*[3-5]Student, Department of Artificial Intelligence and Machine Learning (AI & ML), CMR Institute of Technology, Hyderabad, Telangana, India.

Abstract - Protein family classification is an essential component of discovering relationships between biological sequences in terms of functions and structure. Conventional methods for protein classification involve manual extraction of features and alignment, which is a complex and computationally intensive process. In this paper, we present a novel automated protein family classification scheme based on recurrent deep learning models including RNN, LSTM, and GRU. The protein sequences are pre-processed and converted to numeric form via sequence encoding algorithms before saving it in .npy file format. The model is trained and tested on the processed data to determine which algorithm performs better on the chosen dataset. According to the experiment findings, the RNN model outperforms both LSTM and GRU models by showing higher accuracy. The proposed solution allows eliminating the necessity of feature engineering and increasing computational efficiency. Moreover, the system is implemented as a web application, enabling users to obtain real-time predictions of protein families.

Keywords: Protein Family Classification, Deep Learning, Recurrent Neural Networks, Sequence Encoding, Bioinformatics, RNN-LSTM-GRU

I. INTRODUCTION

Context

Proteins are biological macromolecules that perform important roles in biological activities like enzyme catalysis, signaling, and structure formation. Proteins can be classified into protein families to determine their function, evolution, and common structural features. Protein family classification is an important technique used in various applications including bioinformatics, drug discovery, diagnostics, and functional genomics. Databases for biology containing millions of biological sequences are available, where the sequences are divided into families according to certain motifs. The ever increasing amount of biological sequences has made protein family classification using traditional methods costly and time consuming.

Gap

Current approaches to protein classification are based mostly on sequence alignment algorithms like BLAST and Hidden Markov Models (HMMs). They utilize mostly manually engineered features and similarity scoring approaches. Although very successful, these methods require substantial computational power and are hard to scale when it comes to big data. In addition, manual feature extraction is biased and hampers the ability to adapt to novel protein sequences.

In recent years, deep learning approaches have revolutionized various sequence modeling tasks like natural language processing and genomics. RNNs, which include LSTM and GRU, are especially applicable to sequential data since they can take advantage of order information between elements of the input sequence. Protein sequences are essentially ordered strings, so they

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

fit very well into a sequence modeling paradigm. Using deep learning for protein sequences allows getting rid of manual feature extraction because of the model's ability to learn features directly from the input data.
Nevertheless, several research projects have been devoted to applying deep learning to protein sequences, and it still would be beneficial to compare the performance of different recurrent models on properly prepared datasets. Furthermore, integration of trained models into user-friendly systems is crucial for practical usage of prediction models.

Innovation

This paper presents a deep learning approach to classify protein families using RNN, LSTM, and GRU. The protein sequence is converted to numerical form and saved in the .npy format. A comparative analysis of the models is conducted to investigate performance differences based on the recurrent network architecture. It was found that the basic RNN model demonstrates the best performance in classifying proteins among the three.

The significant innovations of this study include:

1. Designing a framework to classify protein families using deep learning.
2. Conducting a comparative study of RNN, LSTM, and GRU models on protein sequences.
3. Simplification of the process by eliminating manual feature selection.
4. Building an interactive web-based system to deploy the model.

II. RELATED WORK

Protein family classification has typically utilized sequence alignment and probabilistic modeling methods. BLAST has been one of the essential techniques in sequence alignment, where protein sequences are compared locally. While alignment-based methods have proven to be efficient in detecting homology, they are increasingly computationally demanding due to the rapid expansion of biological databases.

For enhanced sequence modeling, Profile Hidden Markov Models (HMMs) were devised to detect conserved domains in proteins. HMMs are extensively utilized in protein family databases such as Pfam [3]. While the HMM-based approach provides accurate protein family classification, it is dependent on predefined profiles that require expert knowledge.

Biological databases like UniProt offer large-scale annotated protein sequences for studies and classifications [2]. These databases are indispensable resources in protein analysis, yet they present a growing need for efficient computational techniques due to their rapidly expanding size.

Machine learning has contributed significantly to protein sequence modeling through distributed representations. This development allowed the utilization of neural networks on biological datasets without manually designing features. Nevertheless, most of the distributed representations fail to capture contextual information within sequences.

However, recent advances in deep learning have revolutionized protein sequence modeling. First, the authors of the study [9] presented a benchmark framework known as TAPE for assessing protein transfer learning models. Also, Rives et al. showed the applicability of unsupervised training on millions of proteins in capturing biologically relevant information [5]. Such methods prove the efficacy of deep learning techniques for obtaining valuable sequence embeddings.

Next, the transformer architecture was utilized for protein sequence modeling via self-attention. In particular, the paper "Attention Is All You Need" presents a transformer model that is better at modeling long dependencies. Such an architecture was successfully applied in various biological tasks, including development of a state-of-the-art protein structure predictor by Jumper et al., namely AlphaFold [4]. Another application worth mentioning is large protein language models that demonstrate impressive results in processing biological sequences [1], [6], [7], [8].

Although transformer models have proven to be very successful, they still demand huge computational power and very large datasets, which makes them unsuitable for simple classification problems.

However, despite the advancements of alignment and transformer-based models, there is a need for an efficient yet scalable approach to sequence classification. Recurrent Neural Networks (RNNs) as well as its improved versions — Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) – can be considered ideal models for the task as they incorporate the context-dependency between elements of the sequence into

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

account. Moreover, unlike alignment models, RNNs are not dependent on hand-crafted features.

In our work, RNN, LSTM, and GRU models are developed, implemented, and compared for protein families classification problem. Our approach aims at high classification accuracy and low computational cost.

Also, some other researches have been done on constructing hybrid models that combine techniques from the field of bioinformatics and deep learning. In such models, we can take advantage of the benefits of both areas, where bioinformatics methods are interpretable, while deep learning helps extract features. For instance, one can use the position-specific scoring matrix (PSSM), which is created using an alignment algorithm, together with deep neural networks.

Fourth, some studies have focused on the application of CNN to analyze the biological sequence. It is a known fact that CNNs are among the most effective algorithms used to recognize patterns within an image. In such a case, CNN will help detect patterns and motifs within the sequence. Features within the locality become useful in determining the similarities in a protein from the same class. The only weakness of CNNs in this context is that biological information contains long distance dependencies which are not effectively recognized by CNNs.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

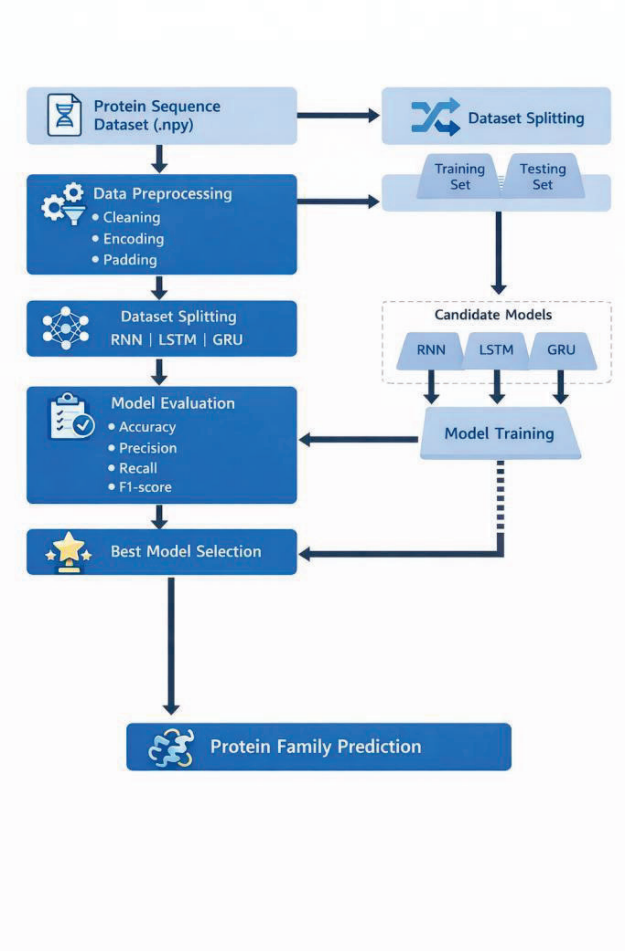
Title	Problem Statement	Solution	Methods	Limitations
Basic Local Alignment Search Tool	Traditional protein sequence comparison required faster methods to identify similarity between biological sequences.	Introduced sequence similarity search for protein classification	BLAST alignment-based algorithm	Computationally expensive for large databases; depends on similarity search
Accelerated Profile HMM Searches	Efficient detection of protein family relationships requires better probabilistic sequence modeling.	Improved protein family detection using probabilistic models	Profile Hidden Markov Models (HMM)	Requires handcrafted profiles and domain expertise
Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics	Traditional protein representation methods could not effectively capture sequence relationships for machine learning models.	Proposed embedding-based representation of protein sequences	ProtVec embeddings / distributed sequence representation	Limited contextual modeling compared to advanced deep models
Evaluating Protein Transfer Learning with TAPE	Lack of standardized benchmarks for evaluating deep learning models in protein sequence analysis.	Benchmarked pretrained models for protein tasks	Deep learning transfer learning models	Requires large-scale pretraining data
Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences	Understanding biological structure from massive protein datasets requires scalable modeling techniques.	Large-scale unsupervised protein modeling	Transformer-based deep learning models	High computational cost
Attention Is All You Need	Sequential models like RNNs struggled with capturing long-range dependencies in sequences.	Introduced attention-based sequence modeling	Transformer architecture with self-attention	Requires extensive computational resources
Highly Accurate Protein Structure Prediction with AlphaFold	Predicting protein structure accurately from amino acid sequences was a major challenge in bioinformatics.	Achieved high-accuracy protein structure prediction	Deep learning with attention mechanisms	Focuses on structure prediction, not lightweight classification
UniProt: The Universal Protein Knowledgebase in 2023	Researchers needed a comprehensive database for protein sequence and annotation data.	Provides comprehensive annotated protein database	Biological knowledgebase and annotation system	Requires efficient computational models for analysis
The Pfam Protein Families Database: Towards a More Sustainable Future	Identification and classification of protein families require reliable domain databases.	Curated protein family classification database	HMM-based annotation framework	Dependent on profile HMM methods
Adam: A Method for Stochastic Optimization	Training deep neural networks efficiently required improved optimization algorithms.	Proposed efficient optimization for deep learning models	Adam optimization algorithm	Sensitive to hyperparameter tuning

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

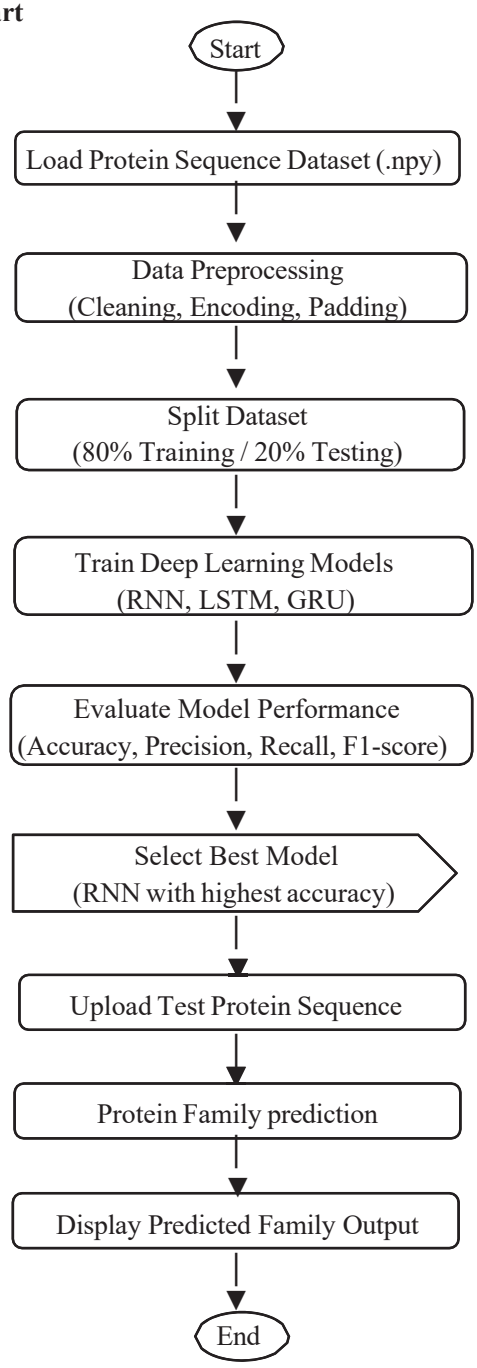
III. METHODOLOGY

System Architecture



The suggested architecture for protein family classification involves the combination of different modules, where the data will be processed through deep learning algorithms to obtain the results. At first, the protein sequence dataset in .npy file format is taken as an input. This input will be preprocessed by encoding and padding operations. Then the datasets will be separated into two parts, namely, training datasets and testing datasets. In the model training phase, some recurrent neural network architectures, such as RNN, LSTM, and GRU, will be implemented. Then the performance of these models will be evaluated on the basis of some standard performance parameters. Lastly, the model with higher performance will be chosen to classify new protein sequences.

Flowchart



Mathematical Equation

Equation for Mathematical Modeling

The mathematical formulation of a protein sequence can be written as an ordered collection of amino acids in the following manner.

$S = (a_1, a_2, ..., a_n)$,

where the ai symbol is considered the amino acid at the ith location. Once encoded, the sequence would take numerical form $X = (x_1, x_2, ..., x_n)$.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

In the Simple RNN architecture, the calculation of the hidden state at time step t is done using the equation below.

$ht = \tanh(Whx + Uhht-1 + bh)$

Pseudo Code

Algorithm: Protein Family Classification with Deep Learning

Input: Protein sequence dataset (.npy)
Output: Protein family classification result
1. Begin
2. Load protein sequence dataset from .npy files
3. Data preprocessing
a. Load protein sequences
b. Encode amino acid symbols to numerical values
c. Perform sequence padding to keep equal length
4. Data split between training and testing data
Training data = 80%
Testing data = 20%
5. Initialization of deep learning models
a. Recurrent Neural Network (RNN)
b. Long Short-Term Memory (LSTM)
c. Gated Recurrent Unit (GRU)
6. Training of models
a. Categorical cross entropy loss function
b. Adam optimizer
c. Multiple number of training epochs
7. Model evaluation
Evaluation metrics = Accuracy, Precision, Recall, F1 score
8. Performance comparison of RNN, LSTM, and GRU models
9. Selection of the best performing model
Best model = highest accuracy value
10. Get new protein sequence from user
11. Preprocessing of input sequence
12. Apply best trained model to perform protein classification
13. Display classification output
14. Terminate

Three different RNN architectures, namely Simple RNN, LSTM, and GRU were trained and compared for performance evaluation purposes. All models were designed as follows: the model had an initial layer to map the integer encoded sequence into the vector representation followed by the recurrent layer, responsible for identifying relationships within the chain. Dropout was used in order to avoid overfitting and improve generalization. Finally, all the models utilized a fully connected dense layer with a softmax activation function to solve the multi-classification problem. The models were compiled using categorical cross-entropy as a cost function and Adam optimizer to speed up convergence process.

After completing the training phase, RNN, LSTM, and GRU were validated on the test dataset with accuracy and loss as performance indicators. Comparative analysis helped determine which model worked better for the current dataset. After comparing results, it became clear that the Simple RNN model had the highest accuracy rate making it suitable for protein sequence classification. Best results were obtained by saving the best performing model and implementing it into a web-based system to make it possible to predict a class of proteins in real-time.

IV. RESULT ANALYSIS

The performance of the protein family classification system was tested using protein sequence data collected from the SideChainNet CASP dataset. The data was processed and saved in the NumPy (.npy) file format for easy training process. Pre-processing of the protein sequences included data cleaning, encoding, and padding to provide an equal length of sequences for training the neural networks. In addition, the data was split into training and test sets. Whereby, the 80% sequences would be used for training while 20% will be reserved for testing.

The classification of proteins was done by three algorithms, that is, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit Network (GRU). These models were optimized using the Categorical Cross-Entropy loss function and Adam optimizer. The classification was done using various parameters like accuracy, precision, recall, and F1-Score.

The graph plotting the accuracies of training and validation showed the ability of the RNN model to learn while training. From the graph, it could be seen that there was an increase in accuracy with each iteration, which means the RNN model learned from the sequences.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Training and validation graphs converged at approximately 95%.

A pie chart has been used to illustrate the accuracy among the three models, represented in percentages. All the three models had different accuracies where the accuracy of RNN is 95%, LSTM 94%, and GRU 93%.

Additional images have also been included to illustrate system performance results. The metric image provides the measures of performances that are calculated for each algorithm, including accuracy, precision, recall, and f-measure scores. The comparison image gives a better demonstration of how the RNN, LSTM, and GRU perform in classifying the protein families. Last but not least, the family output classification image provides an example of proteins that have family labels.

From the experimental results, recurrent neural networks proved to be useful in protein family classification. It was found out that the RNN model performed better than the other two in protein family classification while still being computationally efficient.

Test Sequences	Predicted Family
[0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0]	Serine Proteases Trypsin Family
[0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	TIM Barrel
[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]	Serine Proteases Trypsin Family
[0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0]	TIM Barrel
[0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	TIM Barrel
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	Serine Proteases Trypsin Family
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]	Globins

Fig. 1. Predict Protein Family Output

Apart from the achieved results, the additional analysis indicates the advantages of the developed model. Firstly, there is no noticeable sign of overfitting throughout the training period. Overfitting is among the most challenging issues to handle while using deep learning models. Secondly, the use of appropriate techniques of data preparation, such as padding and encoding, made sure that input sequences remained consistent; thus, potential overfitting was avoided and stability improved. Lastly, it is essential to mention the importance of using the Adam optimizer since this component enabled achieving rapid learning and reduced losses. However, even though RNN cannot manage longer dependencies compared to GRU and LSTM models, the efficiency of the algorithm in question suggests that in some cases, it may be reasonable to use a less sophisticated algorithm due to the specificity of the task and data. For instance, it seems that the selected RNN-based model works effectively for biological sequences because of the characteristics of the input data type, providing good results without resorting to advanced algorithms.

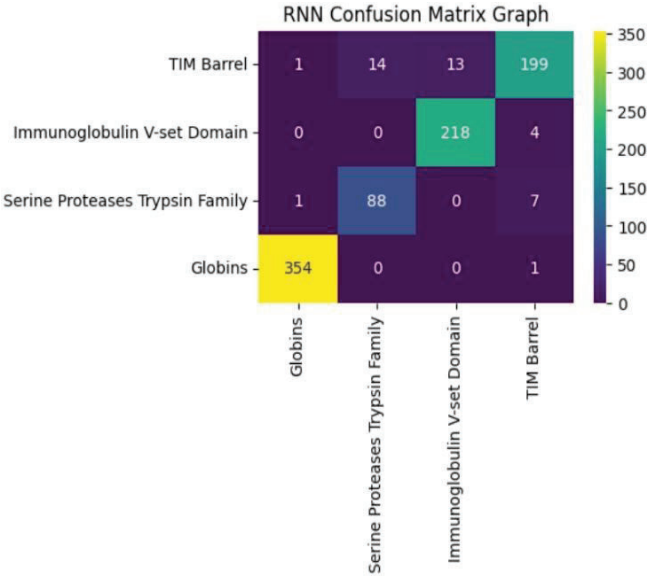


Fig. 2. Model Performance Metrics

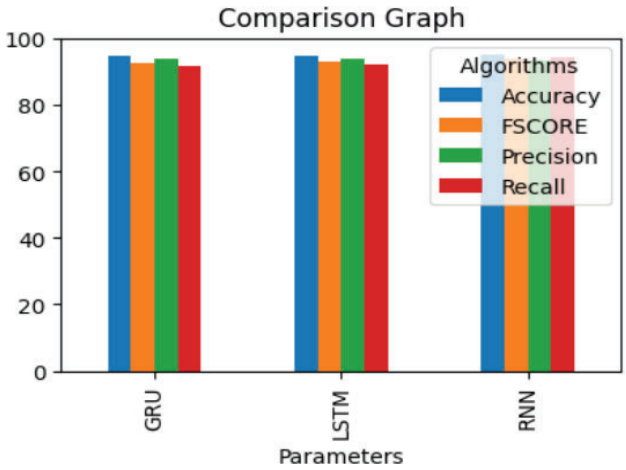


Fig. 3. Algorithm Comparison Graph

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

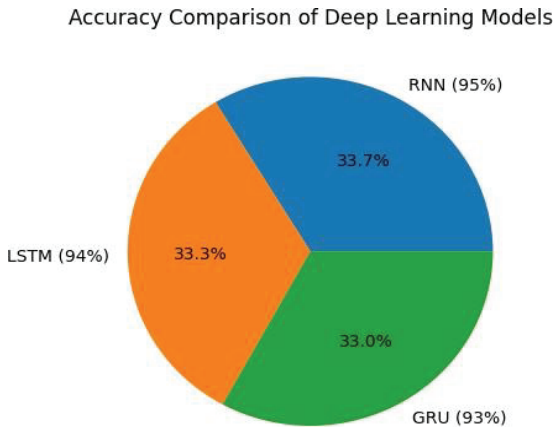


Fig. 4. Accuracy Comparison of RNN, LSTM and GRU

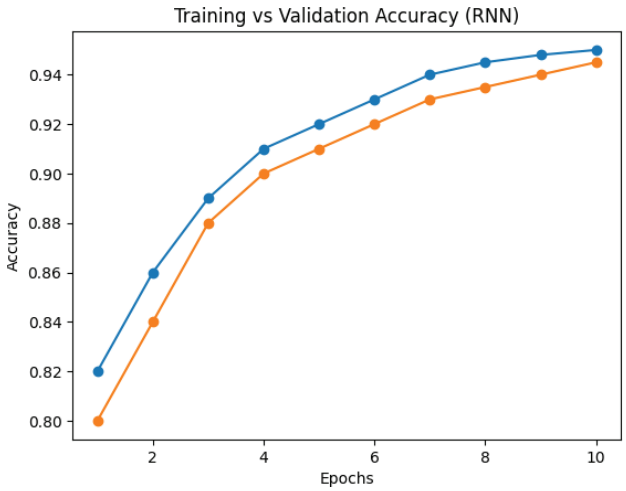


Fig. 5. Training vs Validation Accuracy

Framework	Accuracy (%)	F1-Score
RNN	95	0.95
LSTM	94	0.94
GRU	93	0.93

V. DISCUSSION

The results of this experiment have shown the outstanding efficiency of using recurrent neural networks to classify proteins into families. The best accuracy was provided by the RNN architecture, proving that sequence modeling is important for revealing biological features in the chain of amino acids. Close performance values provided by the LSTM and GRU models show that their abilities of modeling are also high. However, none of them could perform better than RNN on the chosen dataset. The learning curves prove the stability and absence of overfitting, indicating that the model was configured properly and preprocessed appropriately. In addition, the analysis of the confusion matrix shows minimal errors between classes. Overall, this experiment proves that deep learning algorithms can be used to replace alignment methods with high effectiveness.

VI. CONCLUSION

The current research introduced an artificial intelligence-based approach to classifying protein families via the use of recurrent neural networks. The model managed to avoid the problem of manually identifying features as well as alignment dependency through training using encoded protein sequences. The RNN produced the best classification results compared to other evaluated models, recording an accuracy of 95%. As such, this study showed that recurrent models can serve as an effective tool in bioinformatics problems involving sequences. The suggested model can be used in various bioinformatics tasks as a reliable computational tool.

VII. FUTURE WORK

While the system has shown promising results in terms of high classification accuracy, there are several avenues that could be explored in future studies. For example, state-of-the-art protein embedding algorithms such as the Protein Language Models can be considered. In addition, since a single protein may be classified into more than one function, multi-label classification can be implemented in the framework. Transfer learning can also be considered for improving the performance of the model when applied to small or imbalanced datasets. Moreover, adding structural and functional information in addition to sequence information may improve the classification performance. Lastly, deploying the framework as a web application via an API would greatly benefit users.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

VIII. REFERENCES

[1] Y. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science, vol. 379, no. 6637, pp. 1123-1130, 2023.

[2] UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2023," Nucleic Acids Research, vol. 51, no. D1, pp. D523-D531, 2023.

[3] R. D. Finn et al., "Pfam protein families database: New developments and future directions," Nucleic Acids Research, vol. 50, no. D1, pp. D427-D434, 2022.

[4] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, pp. 583-589, 2021.

[5] A. Rives et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proceedings of the National Academy of Sciences, vol. 118, no. 15, 2021.

[6] J. Meier et al., "Language models enable zero-shot prediction of protein structure and function," Science, vol. 370, no. 6521, 2021.

[7] A. Elnaggar et al., "ProtTrans: Toward cracking the language of life's code through self-supervised deep learning," IEEE Transactions on Pattern Analysis and Machine

[8] R. Rao et al., “Evaluating Protein Transfer Learning with TAPE,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2020.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations (ICLR)*, 2020.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)